Power Mean Pyramid Scores for Summarization Evaluation

Sameer Maskey[*]*, Andrew Rosenberg*[+]

[*] IBM Research, New York, NY, USA
[+]Queens College, CUNY, Queens, USA
smaskey@us.ibm.com, andrew@cs.qc.cuny.edu

Abstract

We present Power Mean Pyramid Scores (PMP), an evaluation metric that extends the Pyramid evaluation scheme for summarization by combining Sentence Content Units (SCU) scores using Power Mean. The Pyramid method generates a summarization score by linearly combining component SCU scores. We find that by combining SCU scores using Power Mean, we can optimize a single parameter, α, leading to significantly improved correlation with human judgements. We demonstrate this result through an empirical study based on TAC-08 evaluation.

1. Introduction

Automatic summarization is the task of identifying the most relevant facts in a document or set of documents. There are many existing successful approaches to automatic summarization for newswire text [1], broadcast news speech [2] and video [3]. A challenging problem within summarization is the evaluation of automatically generated summaries. The most popular approaches, including Pyramid Scoring [4], ROUGE [5], and F-measure [6], compare a machine-generated summary against a set of human-generated summaries.

Pyramid Scoring has become a popular technique for summarization evaluation. The idea behind the Pyramid method is that the relevance of a unit of information can be determined by how many reference summaries include it. The unit of information used by the Pyramid method is the Summary Content Unit (SCU). An SCU is a semantically atomic unit representing a single fact, but is not tied to its lexical realization; two paraphrases of the same fact represent the same SCU despite being expressed differently. An SCU is assigned a score proportional to the number of reference summaries that contain it. A Pyramid Score for a summary is calculated by taking a normalized mean of the scores of the contained SCUs. One advantage of Pyramid scores is that it directly assesses the identification of relevant *facts*, while ignoring their *lexical realization*.

In this paper, we describe a method to improve Pyramid Scoring. Power Mean Pyramid Scoring (PMP) uses the original Pyramid method to assign scores to SCUs, but identifies a more reliable combination function than normalized mean. PMP demonstrates significantly higher and more consistent correlation with human judgements of relevance than Pyramid Scores in TAC-08 evaluation framework. One closely related work finds an optimal combination of alignment values for Machine Translation [7]. They show that PM can be effectively used to weight the alignment values such that optimal α produces better combined alignments. Beyond Pyramid and ROUGE, [8] have explored other evaluation measures for summarization. We provide details of Pyramid scores in Section 2 and PMP in Section 3. We evaluate PMP on TAC-08 material (c.f. Section 4), describing the evaluation in Section 5. We conclude and describe future work in Section 6.

2. Pyramid Method for summary evaluation

The "pyramid" in the Pyramid method is composed of SCUs. Each tier corresponds to a number of reference summaries that contain a particular SCU. The motivation behind the name "pyramid" is that the top tier, containing SCUs included by all reference summaries, will have the fewest members, and lower tiers will have successively more.

The first step in the Pyramid method is the identification of all SCUs that appear in any reference summary. Each SCU is then assigned weight proportional to the number of reference summaries that include it. In an evaluation using n reference summaries, the pyramid will have n tiers, indexed by the number of reference summaries including the contained SCUs. To calculate the score, the weight of all included SCUs are added. Let D_i be the number of SCUs in the summary that appear in tier T_i, and let $X = \sum_i D_i$ be the total number of SCUs in the summary. The total SCU weight $D = \sum_{i=1}^{n} i * D_i$. SCUs that do not appear in the pyramid are assigned a weight of zero. This SCU weight is then normalized by the optimal content score for a summary including X SCUs. This guarantees that the Pyramid score will range from zero to one. The optimal content score is calculated

as follows, where $|T_i|$ is the number of SCUs in tier T_i

$$Max = \sum_{i=j+1}^{n} i|T_i| + j(X - \sum_{i=j+1}^{n} |T_i|) \quad (1)$$

where $j = max_i(\sum_{t=i}^{n} |T_i| \geq X)$

Notice that a summary that includes a single SCU from the top tier, T_n, has a pyramid score of 1, regardless of how many optimally relevant SCUs are identified, $|T_n|$. Due to the normalization term, this is an optimal summary score. This normalization makes standard Pyramid scores precision-biased. On the other hand, without any normalization, the maximal pyramid score would be achieved by including all SCUs included by any reference summary. This would result in a recall-biased score, with no penalty for the inclusion of SCUs that occur in no reference summaries.

3. Power Mean Pyramid Scores

The power mean is a generalization of the standard Pythagorean means. The calculation of power mean is shown in Equation 2.

$$M(\alpha, \vec{x}) = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^{\alpha} \right)^{\frac{1}{\alpha}} \quad (2)$$

The power mean formula provides a generalized way to combine a vector of values. At various settings of α, the Pythagorean means, minimum and maximum are special cases of power mean.

$\lim_{\alpha \to -\infty} M_\alpha(x_1, ..., x_n) = min\{x_1, ..., x_n\}$	min
$M_{-1}(x_1, ..., x_n) = \frac{1}{1/x_1 + ... + 1/x_n}$	H
$\lim_{\alpha \to 0} M_\alpha(x_1, ..., x_n) = \sqrt[n]{x_1 x_2 ... x_n}$	G
$M_1(x_1, ..., x_n) = \frac{x_1 + ... + x_n}{n}$	A
$M_2(x_1, ..., x_n) = \sqrt{\frac{x_1^2 + ... + x_n^2}{n}}$	RMS
$\lim_{\alpha \to \infty} M_\alpha(x_1, ..., x_n) = max\{x_1, ..., x_n\}$	max

Figure 1: Power Mean Special Cases

Evaluation scores are often calculated through arithmetic or geometric means of component measures. This is the case for accuracy, F-measure, ROUGE, and Pyramid scores. By identifying an optimal value for α, power mean can be used to identify a combination of measures that has maximal correlation with human judgments of quality, while avoiding arbitrary, intuitive decisions favoring one combination function over another.

In Power Mean Pyramid Scoring (PMP), the power mean function is used to combine SCU scores. Given a summary, we construct $\vec{x}$ where $|\vec{x}|$ is the number of SCUs in the summary and x_i is the SCU weight of the i-th SCU. To be consistent with the original formulation of Pyramid Scores rather than normalizing power mean by n, we normalize these scores by the maximal SCU score attainable with $|\vec{x}|$ SCUs. Under the power mean function, varying α leads to a score that is recall or precision-biased. For example, by setting α to a value that approaches ∞, the combination function becomes the **minimum** SCU weight included in the summary, a precision-biased measure. On the other extreme, setting α to ∞ uses the **maximum** included SCU weight as the final score, a recall-biased score. Similarly, PMP is equivalent to common combination functions at specific settings of α: $\lim_{\alpha \to 0} M(\alpha, \vec{x}) = \sqrt[n]{x_1, ..., x_n}$, Geometric Mean; $M(1, \vec{x}) = \frac{x_1 + ... + x_n}{n}$, Arithmetic Mean; $M(-1, \vec{x}) = \frac{1}{1/x_1 + ... + 1/x_n}$, Harmonic Mean. By calculating the correlation of this score with human judgements we can identify an optimal value of α.

4. Evaluation Material of TAC08

To evaluate PMP, we use the open question answering evaluation material from the 2008 Text Analysis Conference (TAC-08) Opinion Summarization shared task evaluation [9]. These opinion questions asks a system to summarize justifications of a particular point of view. The set includes questions like Why did people enjoy the movie *Good Night and Good Luck*? The TAC-08 data set consists of 22 *targets*. Each *target* has multiple sub-questions on the same topic. Systems were required to generate summaries for each sub-question. Note that scoring and human evaluation was calculated at the *target* level, though systems generated responses at the *question* level. The responses to these question are treated like a traditional summarization task, and are evaluated with a measure based on the Pyramid method.

Participating systems are required to aggregate opinions across blogs and produce an answer containing any support of the proposition. As in a traditional summarization task, the answer may contain many surface realizations of the same supporting facts. By representing the underlying facts rather than the surface forms, the Pyramid method is ideal for this task.

TAC-08 was evaluated using Blog06 [10] dataset containing blog posts, non-blog documents and spam. Each unit of text in the answer of TAC-08 had to be supported from one of the Blog06 documents. For evaluation, ten humans were asked to produce opinion summaries to be used as the reference summaries used in the Pyramid evaluation. Both the reference responses and the system responses were annotated manually for the SCUs. Using these annotations, systems were given a score based on the Pyramid method. In addition to this evaluation, system responses were judged by human annotators on five performance measures, including Overall Responsiveness. To evaluate the merits of the Power Mean Pyramid Scoring, we compare correlations between evaluation measures and the human annotated Overall Respon-

siveness Judgement (ORJ). ORJ was based in a scale of 1 to 10.

To address some of the bias effects of Pyramid Scoring, the TAC-08 evaluation used a combination of F-measure (β=3) and Pyramid Scores [11]. Recall is calculated as the total weight of the SCUs included in a response normalized by the total weight of all SCUs across all reference summaries. Precision is calculated based on the character length of the summary. Formally, let S be the response containing N SCUs, $S_1, \ldots, S_N$ and L be the number of NWS characters in the response. Let the set of reference responses be $R_1, \ldots, R_M \in R$. Finally let $w(S_i)$ be the SCU weight function.

$$P = min(1, \frac{L - N * 100}{L}) \quad (3)$$

$$R = \frac{\sum_{i=1}^{N} w(S_i)}{\sum_{j=1}^{M} w(R_j)} \quad (4)$$

$$TAC08 = \frac{(\beta^2 + 1) * R * P}{\beta^2 P + R} \quad (5)$$

	Pearson ρ	Spearman ρ	Kendall τ
Pyramid	-0.003	0.104*	0.078*
TAC08	0.169*	0.171*	0.128*
PMP (α)	0.206* (0.09)	0.145* (0.5)	0.109* (0.05)

Table 1: Correlations on Tuning Set (*$p < 0.05$)

5. Evaluation and Discussion

Evaluation measures are useful insofar as they emulate human responses or predict task success. For summarization, successful evaluation measures should highly correlate with human judgments of summarization quality. To evaluate their relative efficacy, we compare the correlation between PMP, Pyramid Score and the TAC-08 Metric to the TAC-08 ORJ scores.

We randomly split the data into tune and test set with 11 targets in each set corresponding to 382 summaries for tuning and 378 summaries for testing. We compute Pyramid and TAC-08 scores for all the targets of tuning sets using corresponding human answers. We then calculate correlation of these scores with human ORJ scores. Since there are disagreements between statisticans on which correlation method to use when computing correlation between real numbers and discrete human ratings, we evaluate using three measures: Spearman, Pearson and Kendall. We present the correlation results for the tuning set in Table 1. All of the results are tested for statistical significance with $p < 0.05$. Correlations with human judgments are often calculated and reported with respect to a single summarization system. Pooling multiple system responses leads to lower correlations than previously reported, but allows us to evaluate the robustness of evaluation measures against the performance of multiple systems. Moreover, this strategy allows us to avoid over fitting to specific system idiosyncrasies.

We observe in Table 1 that the TAC-08 metric correlates more strongly than Pyramid Scores with more than 6.64% absolute for Spearman coefficient. This is not surprising, as the TAC-08 metric was constructed specifically for the evaluation of this task. The construction was performed in an ad-hoc fashion, using β of 3 without any validation that this is optimal. PMP presents a mathematically consistent framework to optimize evaluation metrics.

To find an optimal α for PMP metric we performed a linear search on the tuning set using each correlation measure as the optimization function. The tuned PMP scores are significantly higher than Pyramid scores on tuning set as seen in Table 1, but the correlation function was used for optimization. We then apply the optimal α on the held out test set to measure correlation without manual intervention and obtain significantly higher correlation. Results are shown in Table 2. The correlation coefficients

	Pearson ρ	Spearman ρ	Kendall τ
Pyramid	0.169*	0.198*	0.136*
TAC08	0.204*	0.200*	0.143*
PMP (α)	0.231 (0.09)	0.242 (0.5)	0.169 (0.05)

Table 2: Correlations on Test Set (*$p < 0.05$)

obtained on a held out test set were 0.2417 for Spearman, 0.2307 for Pearson and 0.1693 for Kendall respectively. These correlation coefficients are higher than the regular Pyramid scores by 6.17%, 4.4% and 3.29% absolute respectively. All of the scores were statistically significant with with $p < 0.05$. In fact, we observe that tuned PMP scores also show greater correlation than TAC-08 scores on the test set by 2.7%, 4.2% and 2.6% absolute respectively, suggesting that TAC-08 metric was not optimal for TAC-08 evaluation.

The significant improvement with correlation with human scores using PMP on a held out test set allows us to make a few significant observations. First, even though Pyramid itself is very useful metric for summarization, it may be sub-optimal to use the same metric universally in all data sets and annotations. This study finds that evaluation metrics such as Pyramid can be sensitive to combination function. Figure 2 show that optimal α lies somewhere in between the space of geometric mean ($\alpha = 0$) and arithmetic mean ($\alpha = 1$) for Pyramid based summarization evaluation metric. Second, PMP is remarkably stable across tuning and testing sets. This suggests that PMP can represent a reliable evaluation function with relatively little overhead for tuning with an expectation of consistent performance on the test set as well. Third, Power Mean Pyramid Scores delivers a mathematically coherent framework to find the optimal version of Pyra-

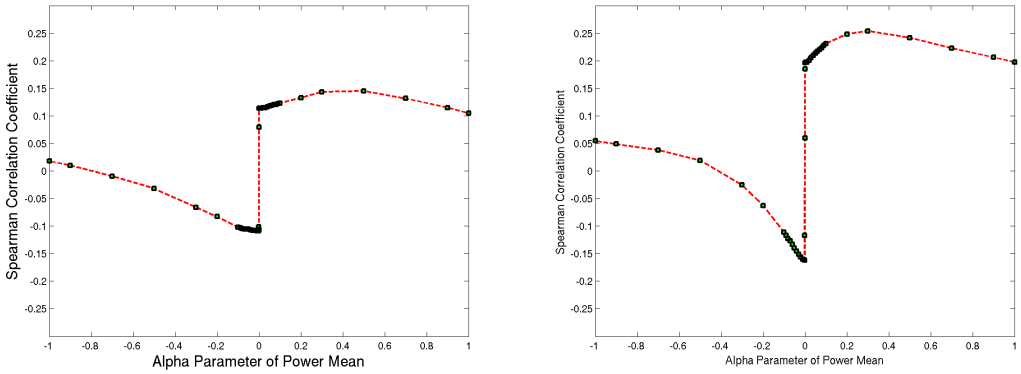

Figure 2: Correlation Graphs with Power Mean under Spearman for Tune and Test Sets

mid metric for the given task, data set and annotation. PMP metric shows higher correlation with human ratings than Pyramid or TAC-08 on the test set (Table 2).

6. Conclusion

We present the Power Mean Pyramid (PMP) Scoring method, a novel mathematical framework that uses Power Mean to compute Pyramid scores. We show that PMP produces summarization scores that have significantly higher correlation with human quality ratings than Pyramid Scores and the TAC-08 metric. We also observe that it may be sub-optimal to use the same metric universally in all data sets and annotations. PMP delivers a mathematically coherent framework to find the optimal version of Pyramid Scores for a given task, data set and annotation. We find that PMP is consistent across tuning and test sets. By identifying a mechanism to optimize evaluation measures with correlation with human judgments, we not only provide the opportunity for improved evaluation on shared-tasks, but provide more informative results to aid in the construction of summarization systems. We believe the generalization of combination functions can be applied to many other evaluation measures including ROUGE and BLEU.

7. References

[1] K. McKeown, R. Barzilay, D. E. John Chen, D. Evans, J. Klavans, A. Nenkova, B. Schiffman, and S. Sigelman, "Columbia's newsblaster: New features and future directions (demo)," in *Proceedings of NAACL-HLT*, 2003.

[2] C. Hori, S. Furui, R. Malkin, H. Yu, and A. Waibel, "Automatic summarization of english broadcast news speech," in *Proc. of the Human Technology Conference (HLT–02), San Diego*, 2002.

[3] L. He, E. Sanocki, A. Gupta, and J. Grudin, "Auto-summarization of audio-video presentations," in *ACM Multimedia (1)*, 1999, pp. 489–498. [Online]. Available: citeseer.ist.psu.edu/he99autosummarization.html

[4] A. Nenkova and R. Passonneau, "Evaluating content selection in summarization: The pyramid method," in *HLT-NAACL 2004: Main Proceedings*, D. M. Susan Dumais and S. Roukos, Eds., Massachusetts, USA, 2004.

[5] C.-Y. Lin, "ROUGE: a package for automatic evaluation of summaries," in *Proc. of workshop on text summarization, ACL-04*, 2004.

[6] C. J. Van Rijsbergen, *Information Retrieval, 2nd edition*. Dept. of Computer Science, University of Glasgow, 1979. [Online]. Available: citeseer.ist.psu.edu/vanrijsbergen79information.html

[7] S. Maskey, S. Rennie, and B. Zhou, "A power mean based algorithm for combining multiple alignment tables," in *Coling 2010*, Beijing, China, August 2010.

[8] H. Jing, R. Barzilay, K. Mckeown, and M. Elhadad, "Summarization evaluation methods: Experiments and analysis," in *In AAAI Symposium on Intelligent Summarization*, 1998, pp. 60–68.

[9] NIST, "Tac 2008 opinion summarization task guidelines," http://www.nist.gov/tac/2008/summarization/op.summ.08.guidelines.html, 2008.

[10] N. NIST, "The blogs06 test collection," http://ir.dcs.gla.ac.uk/test_collections/blog06info.html, 2006.

[11] H. Dang and J. Lin, "Different structures for evaluating answers to complex questions: Pyramids wont topple, and neither will human assessors," in *Association for Computational Linguistics*, 2007, pp. 768–775.